ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio*

KARLHEINZ BRANDENBURG, *AES Fellow*

FhG-IIS, Erlangen, Germany

AND

GERHARD STOLL

Institut für Rundfunktechnik, Munich, Germany

Coauthors:
Yves-François Dehéry, C.C.E.T.T, France
James D. Johnston, AES Member, AT&T, Bell Laboratories, USA
Leon v.d. Kerkhof, Philips, Netherlands
Ernst F. Schröder, AES Member, Thomson Consumer Electronics, Germany

The standardization body ISO/IEC/JTC1/SC29/WG11 (Moving Pictures Expert Group, MPEG) was drafting a standard for compressing the high bit rate of moving pictures and associated audio down to 1.5 Mbit/s. The audio part of the proposed standard is described. Three layers of the audio coding scheme with increasing complexity and performance were defined. These layers were developed in collaboration mainly with AT&T, CCETT, FhG/University of Erlangen, Philips, IRT, and Thomson Consumer Electronics. The generic coding system is suitable for different applications, such as storage on inexpensive storage media or transmission over channels with limited capacity (such as digital audio broadcasting or ISDN audio transmission).

0 INTRODUCTION

The necessity to specify a generic video and audio coding scheme for many applications dealing with digitally coded video and audio and requiring low data rates has led the ISO/IEC standardization body to establish the ISO/IEC JTC1/SC29/WG11, called MPEG (Moving Pictures Experts Group). This group had the task to compare and assess several digital audio low-bit-rate coding techniques in order to develop an international standard for the coded representation of moving pictures, associated audio, and their combination when used for storage and retrieval on digital storage media (DSM). The DSM targeted by MPEG include CD-ROM, DAT, magneto-optical disks, and computer disks, and it is expected that MPEG-based bit-rate reduction techniques will be used in a variety of communication channels such as ISDN and local area networks and in broadcasting applications. The international standard ISO/IEC 11172 "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s" was finalized in November 1992 and consists of three parts: system, video, and audio [1]. The system part (11172-1) deals with synchronization and multiplexing of audio–visual information, whereas the video (11172-2) and audio (11172-3) parts address the video and the audio bit-rate reduction techniques, respectively. This standard is also known as the MPEG-1 standard.

MPEG-2 Audio is the consequent extension from two to five audio channels providing backward compatibility

* Presented at the 92nd Convention of the Audio Engineering Society, Vienna, Austria, 1992 March 24–27; revised 1994 July 15.

to MPEG-1. The main aspects are high quality of five (+1) audio channels, low bit rate and backward compatibility—the key to insuring that existing 2-channel decoders will still be able to decode compatible stereo information from five (+1) multichannel signals.

This standard, which is expected in November 1994, is based on standards and recommendations from international organizations such as ITU-R, SMPTE, and EBU. International standardization bodies will insure the highest audio signal quality by extensive testing. For audio reproduction the loudspeaker positions left, center, right, left and right surround are used, according to the 3/2-standard.

1 STANDARDIZATION AND QUALITY ASSESSMENTS WITHIN MPEG-1 AUDIO

Since 1988 ISO/MPEG has been undertaking the standardization of compression techniques for video and associated audio. The main topic for standardization in MPEG was video coding together with audio coding for DSM. On the other hand the audio coding standard developed by this group was the first international standard in the field of digital audio compression and is expected to be followed in different applications. Beside several subgroups such as video, system, test, implementation, requirement, and DSM, the audio subgroup of MPEG had the responsibility for developing a standard for coding of PCM audio signals with sampling rates of 32, 44.1, and 48 kHz at bit rates in a range of 32–192 kbit/s per mono and 64–384 kbit/s per stereo audio channel. The operating modes are

- Single channel
- Dual channel, like bilingual
- Stereo
- Joint stereo (combined coding of left and right channels of a stereophonic audio program)

Table 1 gives a short general view of the milestones of the MPEG-AUDIO group. This group asked for proposals for the audio coding standard in mid-1989, and 14 proposals were submitted for this purpose. The original proposals were grouped into four clusters according to algorithmic similarities. The clustered candidate algorithms were called ASPEC, ATAC, MUSICAM, and SB/ADPCM.

A number of subjective tests were performed [2]–[4] since mid-1990 to assess the audio quality of the ISO/MPEG/Audio coding standard. During this time period several improvements have been made to meet the present audio quality. The important milestones in the development of the standard have been the official tests organized by the Swedish Broadcasting Corporation in Stockholm under the auspices of ISO and EBU. In July 1990 large listening tests and objective evaluations, such as basic audio quality at different bit rates, sensitivity to transmission bit errors, encoder and decoder complexity, and coding delay, were performed on prototype real-time implementations of the four clustered algorithms. Both the ASPEC and MUSICAM proposals have shown a very high subjective quality at bit rates of about 100 kbit/s per channel.

Due to the result that the proposals of the ASPEC and MUSICAM groups have been subjectively nearly equally rated, and were judged relatively close in their overall performance, the official decision was as follows [5]:

> . . . the MPEG standardization committee decided to approve a collaborative development of the draft audio coding standard between the ASPEC and MUSICAM groups, because the ASPEC codec was slightly superior with respect to the audio quality, especially for lower bit rates (64 kbit/s/channel), and the MUSICAM codec was slightly superior with respect to implementation complexity and decoding delay. The decision was that MUSICAM should be the basis for the low-complexity first layer, and algorithmic refinements including contributions of ASPEC should be used in the subsequent layers.

Table 1. Milestones of ISO/MPEG-Audio group during the development of audio part of the International Standard 11172.

Date	Activities
1988 December	First audio meeting in Hanover
1989 January to 1990 March	Preparation of tests
1989 May	Determining requirements and weighting procedure
1989 June	Proposal of 14 algorithms to be tested
1989 October	Clustering of proponents into four groups
1989 December	Detailed description of four clustered proposals: ASPEC, ATAC, MUSICAM, and SB-ADPCM
1990 May	Exchange of tapes with coded audio sequences between four clusters
1990 June	Subjective and objective tests at SR, Stockholm
1990 August	Presentation of results and decision to follow a layer concept
1990 December	First draft of part 3, "Audio Coding" of International Standard ISO 11172 was prepared.
1991 May	Verification of three layers by subjective testing, again at SR in Stockholm
1991 June	Layers I and II are frozen; Layer III and 'joint stereo coding' are still under discussion
1991 November	Second verification of Layer III and first checking of Joint Stereo Coding by subjective testing at NDR in Hanover
1991 December	Draft of International Standard (DIS) ready for balloting at national standardization bodies
1992 November	International Standard ISO/IEC 11172-3 accepted by national standardization bodies

A three-layer coding algorithm has been defined. These three layers were tested again in April 1991 by the Swedish Broadcasting Corporation [3], and a last verification test for the very low bit rate of 64 kbit/s/channel and "joint stereo coding" was carried out by the University of Hanover under the auspices of NDR in November 1991 [4]. In November 1991 the final proposal, consisting of three modes of operation called "Layers," was adapted by ISO/MPEG [6].

2 BASIC STRUCTURE OF A GENERIC AUDIO CODING SCHEME USING PERCEPTUAL CRITERIA

The basic structure of a perceptual audio coding scheme is shown in Fig. 1.

1) A time–frequency mapping (filter bank) is used to decompose the input signal into subsampled spectral components. Depending on the filter bank used, these are called subband values or frequency lines.

2) The output of this filter bank, or the output of a parallel transform, is used to calculate an estimate of the actual (time-dependent) masking threshold using rules known from psychoacoustics.

3) The subband samples or frequency lines are quantized and coded with the aim of keeping the noise, which is introduced by quantizing, below the masking threshold. Depending on the algorithm, this step is done in very different ways. The complexity varies from block companding to analysis-by-synthesis systems using additional noiseless compression.

4) A frame packing is used to assemble the bit stream, which typically consists of the quantized and coded mapped samples and some side information, such as bit allocation information.

Depending on the focus on either low frequency resolution together with high time resolution or high frequency resolution which leads to only limited time resolution, the systems are usually called subband coders or transform coders.

2.1 Filter Banks

The following list provides a short overview over the most common filter banks used for coding of high-quality audio signals:

1) *QMF-Tree Filter Banks*: Different frequency resolution at different frequencies is possible. Typical QMF-tree filter banks use from 4 to 24 bands. The computational complexity is high.

2) *Polyphase Filter Banks*: These are equally spaced filter banks which combine the filter design flexibility of generalized QMF banks with low computational complexity [7]. It is possible to design the prototype filter in a way that achieves both good frequency resolution (stop-band attenuation better than 96 dB) and good control of possible time-domain artifacts. A polyphase filter bank using 32 bands is used for Layers I and II of the ISO/MPEG audio coder.

3) *DFT, DCT with Sine-Taper Window*: These were the first transforms used in transform coding of audio signals. They implement equally spaced filter banks with 128–512 bands at a low computational complexity. They do not provide critical sampling, that is, the number of time–frequency components is greater than the number of time samples represented by one block length. Another disadvantage of these transforms are possible blocking artifacts.

4) *Modified Discrete Cosine Transform (MDCT, using time-domain aliasing cancellation as proposed in [8])*: This transform combines critical sampling with a good frequency resolution provided by a sine window (compared to a sine-taper window) and the computational efficiency of a fast FFT-like algorithm. Typically 128–512 equally spaced bands are used.

5) *Hybrid Structures* (such as polyphase and MDCT): Using hybrid structures as first proposed in [9] it is possible to combine different frequency resolutions at different frequencies with moderate implementation complexity. A hybrid system consisting of a polyphase filter bank and an MDCT is used in Layer III.

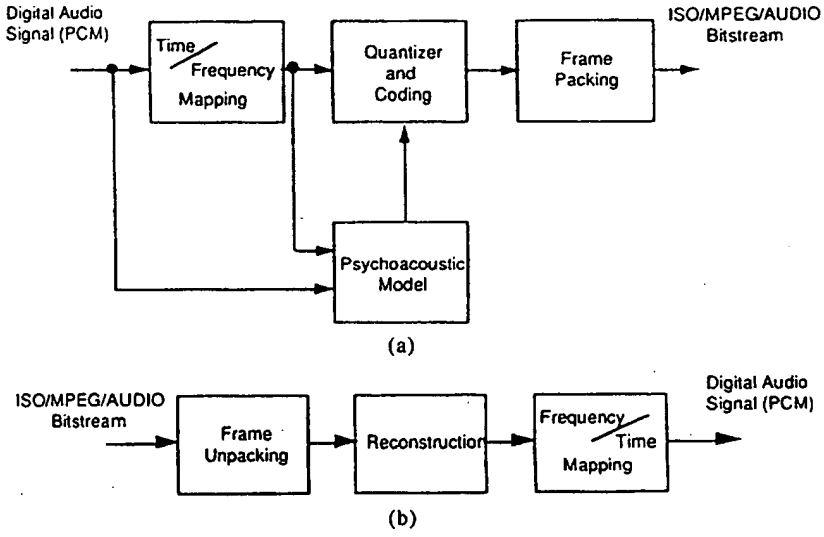

Fig. 1. (a) Basic structure of ISO/MPEG/Audio encoder. (b) Basic structure of ISO/MPEG/Audio decoder.

Theoretically MDCT and polyphase filter banks belong to the same class of time–frequency domain mappings, called lapped orthogonal transform.

3 GENERIC CODING CONCEPT

In view of a number of totally different applications, a concept of a generic coding system was envisioned. Depending on the application, three layers of the coding system with increasing complexity and performance can be used. A standard ISO decoder is able to decode bitstream data which have been encoded in any of the layers. There will also be standard ISO Layer X decoders, which are able to decode Layers X and X $-$ n. The ISO/MPEG/Audio coding technique offers to deal with a much higher dynamic range, due to the scaling technique used, than Compact Disc or DAT, that is, conventional 16-bit PCM.

In all three layers the input PCM audio signal is converted from the time domain into a frequency domain. This is done by a polyphase filter bank consisting of 32 subbands [7].

In Layers I and II a filter bank creates 32 subband representations of the input audio stream, which are then quantized and coded under the control of a psychoacoustic model from which a blockwise adaptive bit allocation is derived.

Layer I is a simplified version of the MUSICAM coding scheme, most appropriate for consumer applications such as digital home recording on tapes, Winchester discs, or on magneto-optical disks, that is, for those applications for which very low data rates are not mandatory.

Layer II introduces further compression with respect to Layer I by redundance and irrelevance removal on the scale factors, and uses more precise quantization. Layer II is nearly identical with the MUSICAM scheme [10], [11], with the exception of the frame header. This header has been added to the MUSICAM frame during the ISO/MPEG/Audio development work. Layer II has numerous applications in both consumer and professional audio, such as audio broadcasting, television, recording, telecommunication, and multimedia [12].

Layer III consists of a combination of the most effective modules of the ASPEC [13] and MUSICAM coding schemes. An additional frequency resolution is provided by the use of a hybrid filter bank. Every subband is thereby further split into higher-resolution frequency lines by a linear transform that operates on 18 subband samples in each subband. In Layer III, nonuniform quantization, adaptive segmentation, and entropy coding of the quantized values are employed for a better coding efficiency. The application of this layer is appropriate most of all in telecommunication, in particular with narrow-band ISDN and in the field of professional audio with high weights on very low bit rates.

Joint stereo coding can be added as an additional feature to any of the layers. This technique exploits the redundancy and irrelevance of typical stereophonic program material and can be used to increase the audio quality at low bit rates or reduce the bit rate for stereophonic signals [14], [15]. The increase of encoder complexity is small and requires negligible additional decoder complexity. Joint stereo coding does not enlarge the overall coding delay.

3.1 Psychoacoustic Models

The psychoacoustic model calculates the minimum masking threshold necessary to determine the just noticeable noise level for each band in the filter bank. The difference between the maximum signal level and the minimum masking threshold is used in the bit or noise allocation to determine the actual quantizer level in each subband for each block. Two psychoacoustic models are given in the informative part of the standard. While they can both be applied to any layer of the MPEG/Audio algorithm, in practice model 1 will be used for Layers I and II, and model 2 for Layer III. In both psychoacoustic models the final output of the model is a signal-to-mask ratio for each subband (Layers I and II) or group of bands (Layer III). The psychoacoustic models are only necessary in the encoder. This allows decoders of significantly less complexity. It is therefore possible to improve even later the performance of the encoder, relating the ratio of bit rate to subjective quality. For some applications which are not demanding a very low bit rate, it is even possible to use a very simple encoder without any psychoacoustic model.

3.1.1 Psychoacoustic Model 1

A high frequency resolution, that is, small subbands in the lower frequency region, and a lower resolution in the higher frequency region with wide subbands should be the basis for an adequate calculation of the masking thresholds in the frequency domain. This would lead to a tree structure of the filter bank. The polyphase filter network used for the subband filtering has a parallel structure which does not provide subbands of different widths. Nevertheless, one major advantage of the filter bank is given by adapting the audio blocks optimally to the requirements of the temporal masking effects and inaudible preechoes. The second major advantage is given by the small delay and complexity. To compensate for the lack of accuracy of the spectrum analysis of the filter bank, a 512-point fast Fourier transform (FFT) for Layer I, and a 1024-point FFT for Layer II are used in parallel to the process of filtering the audio signal into 32 subbands [16]. The output of the FFT is used to determine the relevant tonal, that is, sinusoidal, and nontonal, that is, noise maskers, of the actual audio signal. It is well known from psychoacoustic research that the tonality of a masking component has an influence on the masking threshold. For this reason it is worthwhile to discriminate between tonal and nontonal components. The individual masking threshold for each masker above the absolute masking threshold are calculated depending on frequency position, loudness level, and tonality. All the individual masking thresholds, including the absolute threshold, are added to the so-called global masking threshold. For each subband the minimum

value of this masking curve is determined. Finally the difference between the maximum signal level, calculated by both the scale factors and the power density spectrum of the FFT, and the minimum masking threshold is calculated for each subband and each block. The block-length for Layer I is determined by 12 subband samples, corresponding to 384 input audio PCM samples, and for Layer II by 36 subband samples, corresponding to 1152 input audio PCM samples. This difference of maximum signal level and minimum masking threshold is called signal-to-mask ratio (SMR) and is the relevant input function for the bit allocation.

3.1.2 Psychoacoustic Model 2

The frequency-domain representation of the data is calculated via FFT with a window length of 1024 samples. The calculation is done every 576 samples, that is, synchronous to the hybrid filter bank. The separate calculation of the frequency-domain representation is necessary because the hybrid filter bank values cannot easily be used to get a magnitude–phase representation of the input sequence. The magnitude–phase representation is necessary to calculate the tonality of the current input block for every frequency component.

The tonality estimation works using a simple polynomial predictor, as described in [9]. The basic idea is to use the predictability of the signal as an indicator for tonality. The prediction is done in the magnitude–phase domain. The values stores from the last two blocks are used to predict the magnitude and phase of each frequency line for the current block. The Euclidian distance between estimated and actual values in the magnitude–phase domain is normalized to the maximum possible distance. The normalized value is called "chaos measure" and can assume values between 0 (the rotating phasor prediction had 0 distance from the actual value) and 1 (the predicted value has the maximum distance from the actual value). A logarithmic mapping is used to map the chaos measure range between 0.5 and 0.05 to tonality values of between 0 and 1.

The magnitude values of the frequency-domain representation are converted to a one-third critical band energy representation. A convolution of these values with the cochlea spreading function follows. The next step in the threshold estimation is the calculation of the just masked noise level in the cochlea domain using the tonality index and the convolved spectrum. A correction for the dc gain of the convolution has to be applied. The last step to get the preliminary estimated threshold is the adjustment for the absolute threshold. As the sound pressure level of the final audio output is not known in advance, the absolute threshold is assumed to be some amount below the LSB for the frequencies around 4 kHz. A more detailed description of the estimation of the masking threshold using spreading convolution can be found in [17].

The final step in the calculation of the threshold is preecho control. Preechoes are audible if the backward masking of the signal is not sufficient to mask the error signal, which was spread in time due to the limited time resolution of the synthesis filter bank. This is only possible if there is a sudden increase in signal energy, at least for part of the signal bandwidth. From this a sufficient (but not necessary) condition for the absence of audible preechoes can be derived. The estimated masking threshold is restricted not to exceed the preliminary estimated threshold of the last block. This condition on the final estimated threshold may reduce the estimated threshold by a large amount. To keep the actual quantization noise below this modified threshold, additional bits need to be available to the quantization and coding loop. Layer III contains an intelligent buffer management scheme (called bit reservoir) in order to make the additional bits available when needed. This technique was taken from OCF (see [18]).

4 LAYER I AND LAYER II CODING SCHEME

Block diagrams of the Layer I and Layer II encoders are given in Fig. 2. The coding technique for these layers is based on a subband splitting of the input PCM audio signal by a polyphase analysis filter bank into 32 equally spaced subbands, a dynamic bit allocation derived from a psychoacoustic model, block companding of the subband samples, and the bit-stream formatting [10], [11], [19]. The individual steps of the encoding and decoding process are explained in detailed form in the following sections.

4.1 Filter Bank

The prototype QMF filter is of order 511. It is optimized in terms of spectral resolution and rejection of side lobes, which is better than 96 dB. This rejection is necessary for a sufficient cancellation of aliasing distortions. This filter bank provides a reasonable tradeoff between temporal behavior on one side and spectral accuracy on the other. A time–frequency mapping providing a high number of subbands facilitates the bit-rate reduction due to the fact that the human ear perceives the audio information in the spectral domain with a resolution corresponding to the critical bands of the ear, or even lower. These critical bands have a width of about 100 Hz in the low-frequency region, that is, below 500 Hz, and widths of about 20% of the center frequency at higher frequencies. The requirement of having a good spectral resolution is unfortunately contradictory to the necessity of keeping the transient impulse response, the so-called pre- and postecho, within certain limits in terms of temporal position and amplitude compared to the attack of a percussive sound. The knowledge of the temporal masking behavior [20] gives an indication of the necessary temporal position and amplitude of the preecho generated by a time–frequency mapping in such a way that this preecho, which normally is much more critical compared to the postecho, is masked by the original attack. In conjunction with the dual-synthesis filter bank located in the decoder, this filter technique provides a global transfer function optimized in terms of perfect impulse response perception.

In the decoder the dual-synthesis filter bank recon-

structs a block of 32 output samples. The filter structure is extremely efficient for implementation in a low-complexity and non-DSP-based decoder and requires generally fewer than 80 integer multiplications or additions per PCM output sample. Moreover, the complete analysis and synthesis filter gives an overall time delay of only 10.5 ms at a 48-kHz sampling rate.

4.2 Determination and Coding of Scale Factors

The calculation of the scale factor for each subband is performed for a block of 12 subband samples. The maximum of the absolute value of these 12 samples is determined and quantized with a word length of 6 bits, covering an overall dynamic range of 120 dB per subband with a resolution of 2 dB per scale factor class. In Layer I a scale factor is transmitted for each block and each subband that has no 0-bit allocation.

Layer II uses an additional coding to reduce the transmission rate for the scale factors. Due to the fact that in Layer II a frame corresponds to 36 subband samples, that is, three times the length of a Layer I frame (see Fig. 3), three scale factors have to be transmitted in principle. To reduce the bit rate for the scale factors, a coding strategy which exploits the temporal masking effects of the ear has been studied. Three successive scale factors of each subband of one frame are considered together and classified into certain scale factor patterns. Depending on the pattern, one, two, or three scale factors are transmitted with an additional scale factor select information consisting of 2 bits per subband. If

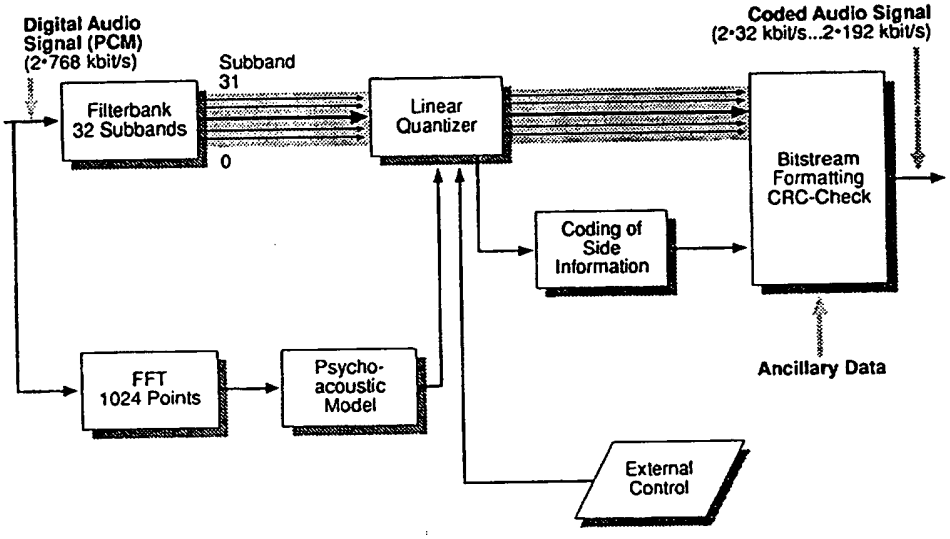

Fig. 2. Block diagram of ISO/MPEG/Audio encoder, Layer I and II (single-channel mode).

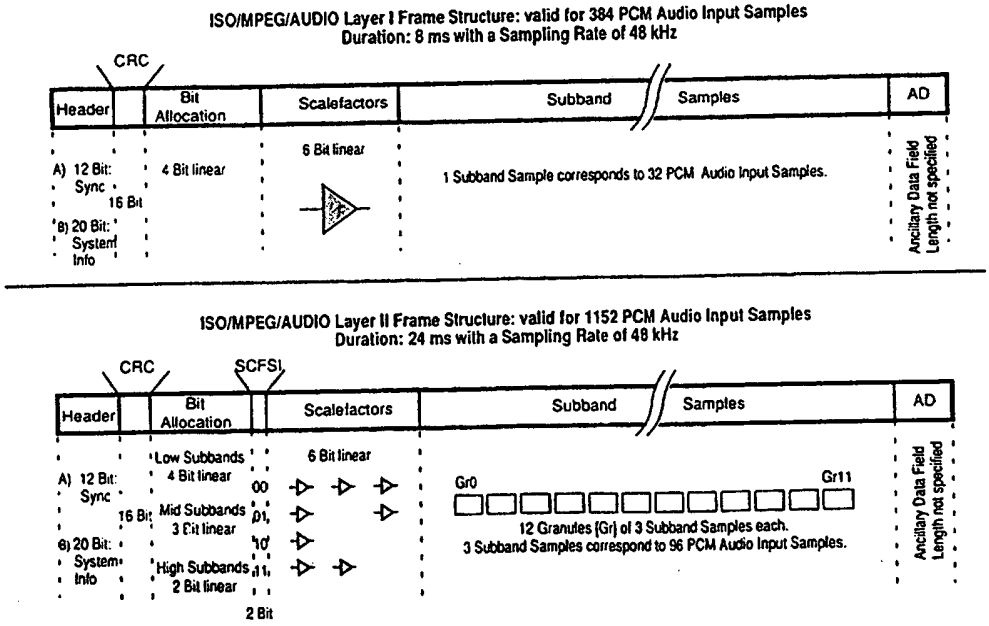

Fig. 3. ISO/MPEG/Audio Layer I and II bit stream.

there are only small deviations from one scale factor to the next, only the bigger one has to be transmitted. This occurs relatively often for stationary tonal sounds. If attacks of percussive sounds have to be coded, two or all three scale factors have to be transmitted, depending on the rising and falling edge of the attack. This additional coding technique allows on average a factor of 2 reduction in the bit rate for the scale factors compared with Layer I.

4.3 Bit Allocation and Encoding of Bit Allocation Information

Before the adjustment to a fixed bit rate, the number of bits available for coding the samples must be determined. This number depends on the number of bits required for scale factors, scale factor select information, bit allocation information, and ancillary data.

The bit allocation procedure is determined by minimizing the total noise-to-mask ratio over every subband and the entire frame. This procedure is an iterative process where in each iteration step the number of quantizing levels of the subband that has the greatest benefit is increased with the constraint that the number of bits used must not exceed the number of bits available for that frame. Layer I uses 4 bits for coding the bit allocation information for each subband and frame, whereas Layer II uses 4 bits for the lowest subbands only and 2 bits for the highest.

4.4 Quantization and Encoding of Subband Samples

First each of the 12 subband samples of one block is normalized by dividing its value by the scale factor. The result is quantized according to the number of bits spent by the bit allocation block. Only odd numbers of quantization levels are possible, allowing an exact representation of a digital zero. Layer I uses 14 different quantization classes, containing $2^n - 1$ steps, with $2 \leq n \leq 15$ different quantization levels. This is the same for all subbands. In addition no quantization at all can be used if no bits are allocated to a subband. In Layer I each sample is coded independently by one code word.

In Layer II the number of different quantization levels depends on the subband number, but the range of the quantization levels always covers a range of 3 to 65 535 with the additional possibility of no quantization at all. Samples of subbands in the low-frequency region can be quantized with 15, in the midfrequency range with 7, and in the high-frequency range with only 3 different quantization levels. The classes may contain 3, 5, 7, 9, 15, 63, . . . , 65 535 quantization levels. Since 3, 5, and 9 quantization levels do not allow an efficient use of a code word consisting only of 2, 3, or 4 bits, three successive subband samples are grouped together to a "granule." Then the granule is coded with one code word. The coding gain by using the grouping is up to 37.5%. Due to the fact that many subbands, especially in the high-frequency region, are typically quantized with only 3, 5, 7, and 9 quantization levels, the reduction factor of the length of a code word is considerable.

4.5 Layer I and Layer II Bit-Stream Structure

The bit stream of these layers was constructed in such a way that both a decoder of low complexity and low decoding delay can be used, and that the encoded audio signal contains a number of entry points with short and constant time intervals. The encoded digital representation of an efficient coding algorithm specially suited for storage application must allow multiples of entry points in the encoded data stream to record, play, and edit short audio sequences and to define the editing positions precisely. To enable a simple implementation of the decoder, the frame between those entry points must contain the whole information which is necessary for decoding the bit stream. Due to the different applications such a frame has to carry in addition all the information necessary for allowing a large coding range with many different parameters. These features are important too in the field of digital audio broadcasting, where a low-complexity decoder is necessary for economical reasons and where frequent entry points in the bit stream are needed, allowing an easy block concealment of consecutive erroneous samples impaired by burst errors.

The format of the encoded audio bit stream for Layers I and II is shown in Fig. 3. The structure of the bit stream is characterized by short autonomous frames corresponding to either 384 PCM samples (8 ms for Layer I at 48-kHz sampling rate) or 1152 PCM samples (24 ms for Layer II at 48 kHz).

4.6 Layer I and Layer II Decoding

The block diagram of the decoder is shown in Fig. 4. First of all, the header information, CRC check, side information, that is, the bit allocation information with scale factors, and 12 successive samples of each subband signal are separated from the ISO/MPEG/Audio Layer I and II bit stream. The reconstruction process to obtain again PCM audio is characterized by filling up the data format of the subband samples with regard to the scale factor and bit allocation for each subband and frame. The synthesis filter bank reconstructs the complete broad-band audio signal with a bandwidth of up to 24 kHz. The decoding process needs significantly less computation power than the encoding process. The relation for Layer I is about 1:2 and for Layer II it is even 1:3. Due to the low computation power needed and the straightforward structure of the algorithm, both layers can be easily implemented into a special VLSI. Since 1993, stereo decoder chips have been available from several manufacturers. Layer I and II stereo encoders are available which are implemented in only one fixed-point DSP (DSP56002).

5 LAYER III CODING SCHEME

A block diagram of the Layer III encoder is shown in Fig. 5. The corresponding decoder is shown in Fig. 6. The filter bank used in Layer III is a switched hybrid polyphase/MDCT filter bank. In the implementation tested by ISO the psychoacoustic model 2 is used to estimate the masking threshold. Nonuniform quantiza-

tion and Huffman coding are used to increase the coding efficiency. A buffer technique called bit reservoir is used to maintain coding efficiency and to keep the quantization noise below the masking threshold. Some of the blocks are explained in more detail in this section.

5.1 Polyphase/MDCT Hybrid Filter Bank

This filter bank was designed with the aim of offering compatibility with Layer II combined with advanced features. The output samples of each of the 32 channels

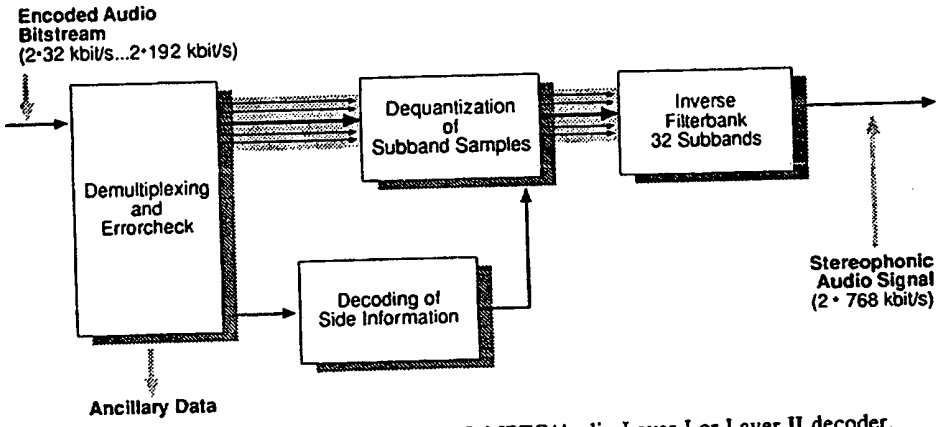

Fig. 4. Block diagram of two-channel ISO/MPEG/Audio Layer I or Layer II decoder.

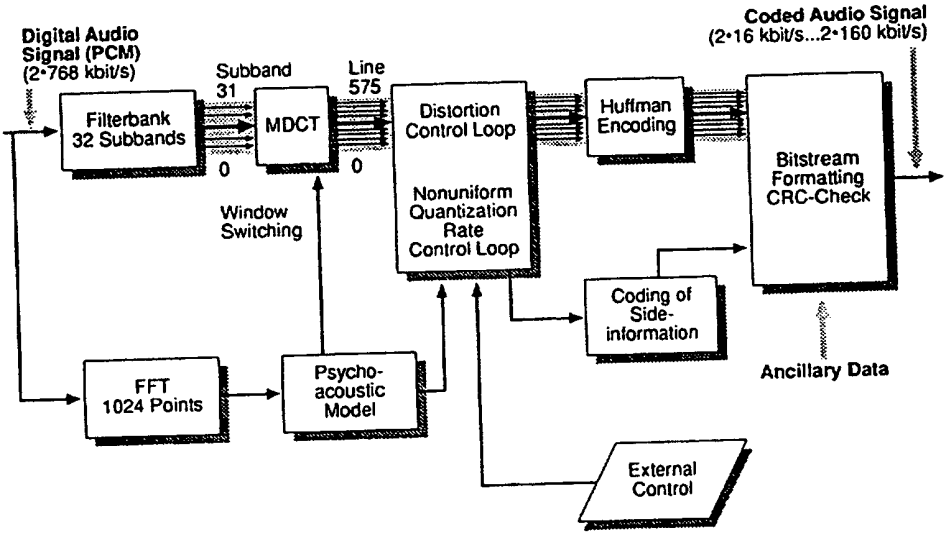

Fig. 5. Block diagram of ISO/MPEG/Audio encoder, Layer III (single-channel mode).

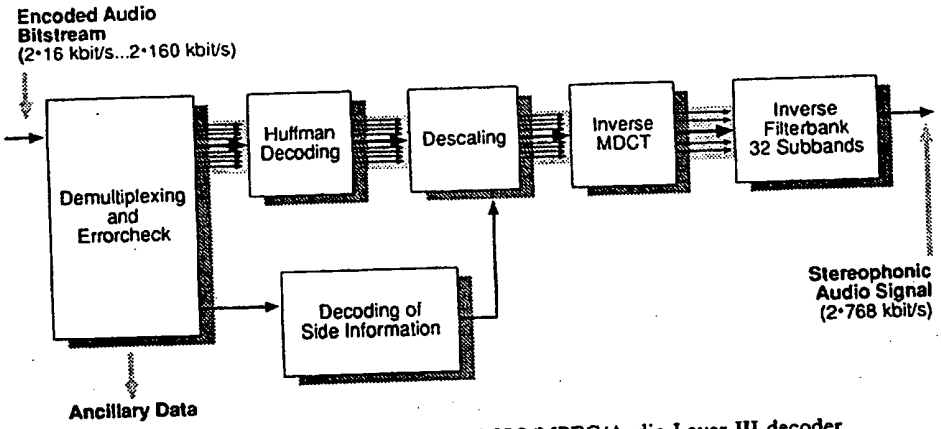

Fig. 6. Block diagram of two-channel ISO/MPEG/Audio Layer III decoder.

of the polyphase filter bank used by Layer II are fed into an 18-channel MDCT filter bank. The maximum total number of output channels is $32 \times 18 = 576$. Due to the small number of channels a direct (matrix multiplication) implementation of the filter bank can be used without much penalty in terms of complexity.

The idea of a hybrid filter bank was first used in [9]. It provided different time–frequency resolutions at different frequencies in order to simulate the frequency–time resolution of the human auditory system. In Layer III the resolution is normally kept constant throughout the spectrum. This leads to the maximum transform gain for stationary signals. If necessary, a part or all of the MDCT filter banks can be switched to lower frequency resolution and higher time resolution.

The window length is 36 in the case of long windows (overlap factor of 2 used in the MDCT) and 12 in the case of short windows. In order to maintain the time-domain alias cancellation property of the MDCT the number of lines must be a multiple of 4. This is the reason why 3:1 switching is used.

5.2 Window Switching and Buffer Control

A window length of 1152 samples corresponds to 24 ms at 48-kHz sampling frequency. All quantization errors in the frequency domain are spread over this time length. For signals containing attacks or similar time-domain events (triangle, castanets) this results in audible preechoes (see [21]). One method used to avoid preechoes is based on the possibility of dynamic changes in the window shape. This technique is based on the fact that alias terms which are caused by subsampling in the frequency domain of the MDCT are constrained to either half of the window. Adaptive window switching as used in Layer III is based on [22].

Fig. 7 shows the different windows used in Layer III, and Fig. 8 shows a typical sequence of window types if adaptive window switching is used. The function of the different window types is explained as follows:

1) *Long Window*. This is the normal window type used for stationary signals.

2) *Start Window*. In order to switch between the long and short window types, this hybrid window is used. The left half has the same form as the left half of the long window type. The right half has the value of 1 for one-third of the length and the shape of the right half of a short window for one-third of the length. The remaining one-third of the window is 0. Thus alias cancellation can be obtained for the part that overlaps the short window.

3) *Short Window*. The short window has basically the same form as the long window, but with one-third of the window length. It is followed by a MDCT of one-third length. The time resolution is enhanced to 4 ms at 48-kHz sampling frequency. The combined frequency resolution of the hybrid filter bank in the case of short windows is 192 lines, compared to 576 lines for the normal windows used in Layer III.

4) *Stop Window*. This window type enables the switching from short windows back to normal windows.

A criterion when to switch the window form is necessary to get the preecho control working.

The use of a hybrid filter bank facilitates advanced preecho control schemes. In the case of a preecho condition all or part of the MDCT filter banks are switched to a better time resolution, as described. The criterion to switch the filter bank is derived from the threshold calculation. If preecho control is implemented in the threshold calculation as described, preecho conditions result in a much increased estimated perceptual entropy (PE) [17], that is, in the amount of bits needed to encode the signal. If the demand for bits exceeds the average value by some extent, a preecho condition is assumed and the window switching logic activated. Experimental data suggest that a big surge in PE is always due to preecho conditions. Therefore preecho detection via the threshold calculation works without errors.

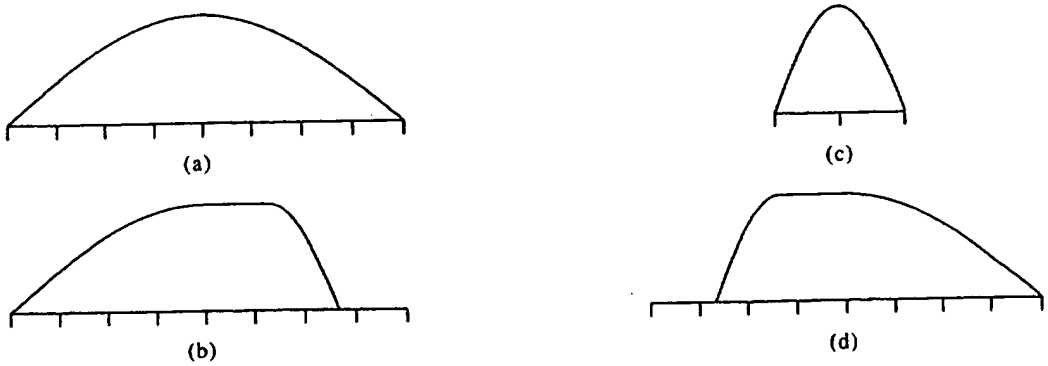

Fig. 7. Different types of windows used for MDCT in ISO/MPEG/Audio Layer III. (a) Normal window. (b) Start window. (c) Short window. (d) Stop window.

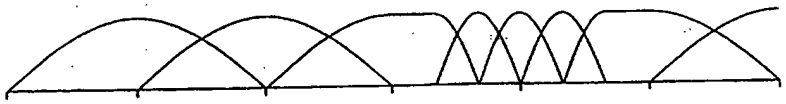

Fig. 8. Typical sequence of window types for adaptive window switching in ISO/MPEG/Audio Layer III.

5.3 Quantization

Layer III uses nonuniform quantization. The basic formula is

$$is(i) = nint\ (((xr(i)/quant)^{**}0.75) - 0.0946)$$

where

xr(i) = absolute value of frequency line at index i
quant = actual quantizer step size
nint = nearest integer function
is(i) = quantized absolute value at index i

The quantization is of the midtread type, that is, values around zero get quantized to zero and the quantizer is symmetric.

The nonuniform quantization was chosen to implement some noise shaping by default. Bigger values are quantized less accurately than smaller values.

5.4 Huffman Coding

The quantized information is coded using several different coding methods. A series of zero at high frequencies is coded by run length coding. For the next region with values not exceeding 1 in magnitude a four-dimensional Huffman code is applied. The remaining "big values" region will be coded with a two-dimensional Huffman scheme and can optionally be split into three subregions, each having a separately selectable Huffman code table. By individually adapting code tables to subregions coding efficiency is enhanced, and simultaneously sensitivity against transmission errors is decreased. The largest tables used for Layer III contain 16 by 16 entries. Larger values are coded using an escape mechanism.

5.5 Analysis by Synthesis

The frequency lines are quantized and coded within two nested loops. In the first loop the overall quantizer step size is adjusted to ensure that the amount of data needed for coding the information does not exceed the number of bits available for one block.

In the second (outer) loop the calculated solution is evaluated with respect to the psychoacoustic demands imposed by masking conditions. This is done in an analysis-by-synthesis procedure, which compares the actual quantization error to the previously calculated masking threshold and accordingly adapts the individual weighting factor of each scale-factor band.

5.6 Bit-Stream Structure

The bit-stream organization closely follows Layer II. The frame length for Layers III and II is identical. Each frame of 1152 time-domain samples is subdivided into two granules of 576 samples. The frame header (as used for all ISO/MPEG/Audio layers) is followed by the side information common to all granules. The side information blocks for the granules follow. They are of constant length (59 bits each) through all modes. The length of the main information for each granule is explicitly contained as part of the side information. This makes it very easy to address the ancillary information, which is located last in every block. The length of the total side information as well as the length of the main information are always an integer number of bytes.

5.7 The Bit Reservoir Technique

The window switching as described does not succeed in avoiding all audible preechoes. This is due to the fact that even the length of the "short" windows corresponds to a time of 8 ms at 48-kHz sampling frequency. Due to this fact a combination of the PE-driven preecho control, as described in Section 3.1 and the window switching is used to prevent audible preechoes. Unfortunately there is a large increase in PE, that is, in the amount of bits needed to code the signal for the frames where preecho control is used.

A buffer technique called bit reservoir was introduced to satisfy the additional need for bits due to the preecho control. It can be described as follows. The amount of bits corresponding to a frame is no longer constant, but varies with a constant long-term average. To accomodate fixed-rate channels, a maximum accumulated deviation of the actual bit rate to the target (mean) bit rate is allowed. The deviation is always negative, that is, the actual mean bit rate is never allowed to exceed the channel capacity. An additional delay in the decoder takes care of the maximum accumulated deviation from the target bit rate.

If the actual accumulated deviation from the target bit rate is zero, then (by definition) it holds that the actual bit rate equals the target bit rate. In this case the bit reservoir is called empty. If there is an accumulated deviation of n bits, then the next frame may use up to n bits more than the average number without exceeding the mean bit rate. In this case the bit reservoir is said to hold n bits.

This is used in the following way in Layer III. Normally the bit reservoir is kept at somewhat below the maximum number (accumulated deviation). If there is a surge in PE due to the preecho control, then additional bits "taken from the reservoir" are used to code this particular frame according to the PE demand. In the next few frames every frame is coded using some bits less than the average amount. The bit reservoir gets "filled up" again.

Figs. 9 and 10 show examples of the succession of frames with different amounts of bits actually used. A pointer called "main data begin" is used to transmit the information about the actual accumulated deviation from the mean bit rate to the decoder. The side information is still transmitted with the frame rate as derived from the channel capacity (mean rate). The main data begin pointer is used to find the main information in the input buffer of the decoder.

6 APPLICATIONS OF ISO/MPEG/AUDIO

There is a wide field of applications for low-bit-rate audio coding schemes. The applications are in the areas of consumer audio as well as professional audio. Any

medium with a channel capacity of 256 kbit/s can be used to distribute a stereophonic audio program with either Layer II or Layer III with no subjective degradation compared to 16-bit PCM. If the channel capacity allows a bit rate of 384 kbit/s, even the low-complexity Layer I can be used for highest audio quality. With bit rates of 64 kbit/s per channel or 2*64 kbit/s per stereo program MPEG-Audio Layer II and even more Layer III provide a subjective audio quality that comes close to the original 16-bit PCM for normal audio material. In 1992 and 93, the specialists group TG 10/2 of ITU-R performed a lot of tests with different types of codecs for applications such as contribution, distribution, commentary and emission. Some of the tests included a cascade of up to nine codecs. The tests showed that only the MPEG-AUDIO Layers fulfilled the requirements of ITU-R. Layer II received the recommendation for contribution, distribution and emission, and Layer III was recommended for commentary application.

The first applications of MPEG-Audio are the following:

- Storage and editing of digital audio on small computers (home studio).
- Computer-based multimedia.
- Digital audio broadcasting (terrestrial and satellite). DAB which was developed by the EUREKA 147 project, and ADR (ASTRA Digital Radio) are using Layer II for sound coding.
- Multichannel audio for ADTV and HDTV.
- Audio recording and reproduction with magnetic tapes, Winchester disks, magneto-optical disks, or solid-state storage media.
- Transmission via narrow-band ISDN for reporting links and tele- or videoconferencing.

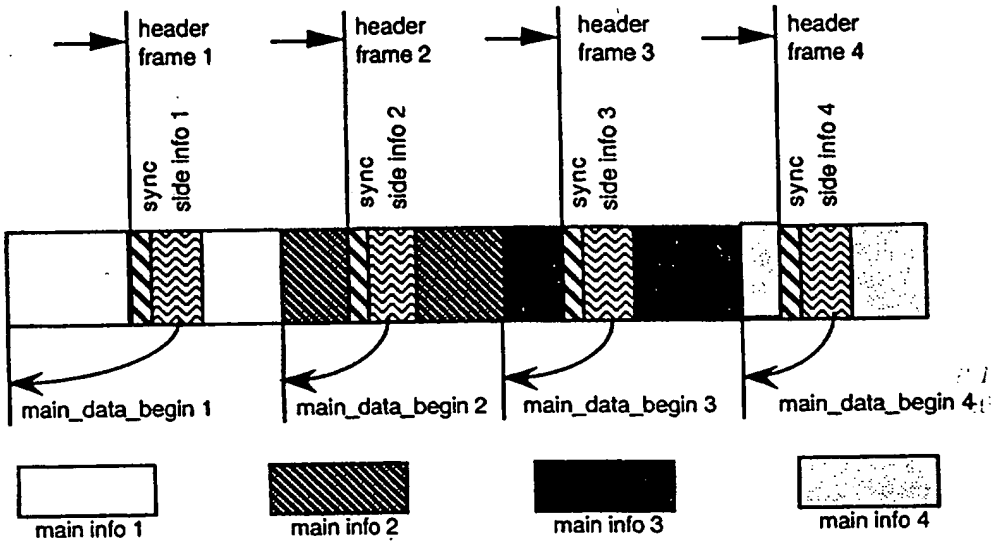

Fig. 9. Layer III bit-stream organization.

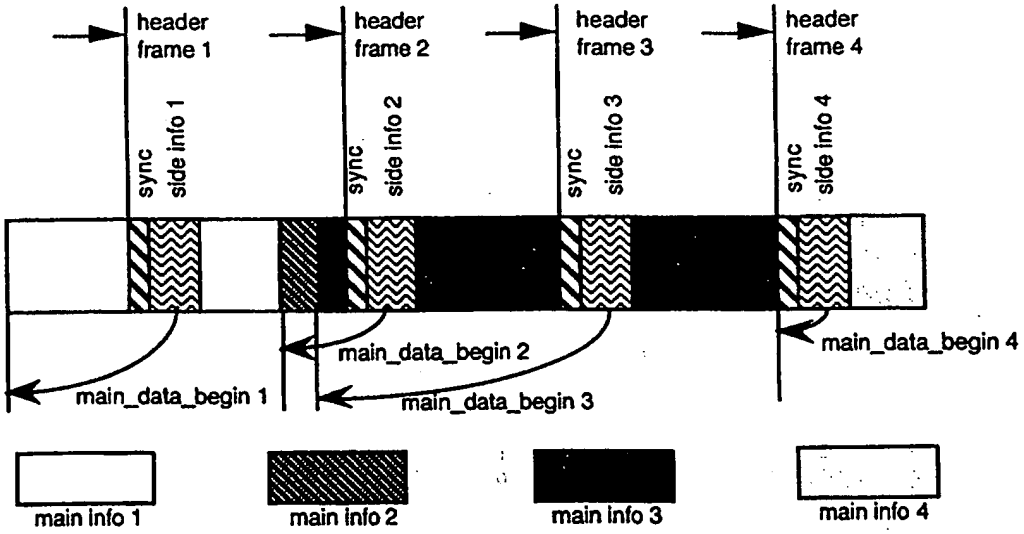

Fig. 10. Layer III bit-stream organization with peak demand at main information 3 and small demand at main information 2.

- Distribution from the studio to transmitter stations and contribution links between studios via ISDN.

7 CONCLUSIONS

In the last 10 years high-quality audio coding went from the first research project to the first commercial applications. With MPEG-1 Audio, the first phase of the development of high-quality audio coding is finished.

Perceptual coders using either high-time-resolution polyphase filter banks or high-frequency-resolution transform filter banks have reached the quality necessary for widespread use in broadcasting, telecommunication, computer, and consumer applications.

The new MPEG-1 Audio coder delivers state-of-the-art performance. A range of coding modes is provided ranging from Layer I, which allows a very simple decoder implementation, to Layer III, which delivers the best performance at 64-kbit/s per audio channel.

8 ACKNOWLEDGMENT

The authors wish to thank their partners, in particular their European partners at the EUREKA 147 project, who have greatly contributed to the audio part of the ISO Standard 11172. Special thanks are due to Nikil Jayant at AT&T Bell Laboratories, Heinz Gerhäuser, Ernst Eberlein, Bernhard Grill, Jürgen Herre, and Harald Popp at FhG-IIS, Michel Lever, Jean-Yves Roudaut, and Pierre Urcun at CCETT, Günther Theile, Martin Link, Detlef Wiese, Robert Sedlmeyer, and Andy Brefort at IRT, Raymond Veldhuis, Frank Zijderveld, Paul Dillen, and Paul De Wit at Philips, and Jens Spille and W. Voessing at Deutsche Thomson Brandt, for their invaluable contributions.

9 REFERENCES

[1] ISO/IEC IS11172-3, "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s—Audio Part" (1992 Nov.).

[2] ISO/IEC JTC1/SC2/WG11 N0030, "MPEG/AUDIO Test Report," Stockholm (1990 July).

[3] ISO/IEC JTC1/SC2/WG11 MPEG 91/010, "SR-Report on the MPEG/AUDIO Subjective Listening Test," Stockholm (1991 April/May).

[4] ISO/IEC JTC1/SC2/WG11 MPEG 91/331, "Report on the MPEG/AUDIO Subjective Listening Tests," Hanover (1991 Nov.).

[5] L. Chiariglione, D. Le Gall, H. G. Musmann, and A. Simon, "Status Report of ISO/MPEG," Doc. ISO/IEC JTC1/SC2/WG11 MPEG90/263 (1990 Sept.).

[6] L. Chiariglione et al., "Press Release of ISO/MPEG," Doc. ISO/IEC JTC1/SC2/WG11 MPEG91/346 (1991 Nov.).

[7] J. H. Rothweiler, "Polyphase Quadrature Filters—A New Subband Coding Technique," in *Proc. Int. Conf. IEEE ASSP*. (Boston, MA, 1983), pp. 1280–1283.

[8] J. Princen, A. Johnson, and A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," in *Proc. ICASSP 1987*, pp. 2161–2164.

[9] K. Brandenburg and J. D. Johnston, "Second Generation Perceptual Coding: The Hybrid Coder," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 38, p. 383 (1990 May), preprint 2937.

[10] Y. F. Dehéry, G. Stoll, and L.v.d. Kerkof, "MUSICAM Source Coding for Digital Sound," in *Symp. Rec. "Broadcast Sessions" of the 17th Int. Television Symp.* (Montreux, Switzerland, 1991 June), pp. 612–617.

[11] G. Stoll, D. Wiese, und M. Link: *MUSICAM: Ein Quellencodierverfahren zur Datenreduktion hochqualitativer Audiosignale für universelle Anwendung im Bereich der digitalen Tonübertragung und -speicherung*, Taschenbuch der Telekom Praxis (Fachverlag Schiele & Schön, Berlin, 1991), pp. 96–127.

[12] G. Stoll and Y. F. Dehéry, "High Quality Audio Bit-Rate Reduction System Family for Different Applications," in *Proc. IEEE ICC'90 Supercom*, vol. 3, no. 322.3 (Atlanta, GA, 1990), pp. 937–941.

[13] K. Brandenburg, J. Herre, J. D. Johnston, Y. Mahieux, and E. F. Schroeder, "ASPEC—Adaptive Spectral Perceptual Entropy Coding of High Quality Music Signals," presented at the 90th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 380 (1991 May), preprint 3011.

[14] R. G. van der Waal and R. N. J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals," in *Proc. ICASSP 1991*.

[15] J. D. Johnston, "Perceptual Transform Coding of Wideband Stereo Signals," in *Proc. ICASSP 1990*.

[16] D. Wiese and G. Stoll, "Bitrate Reduction of High Quality Audio Signals by Modeling the Ears' Masking Thresholds," presented at the 89th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 38, p. 872 (1990 Nov.), preprint 2970.

[17] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Commun.*, vol. 6, pp. 314–323 (1988).

[18] K. Brandenburg, "Ein Beitrag zu den Verfahren und der Qualitätsbeurteilung für hochwertige Musikcodierung," Ph.D. Thesis, Erlangen, Germany, 1989.

[19] G. Theile, G. Stoll, and M. Link, "Low Bit-Rate Coding of High-Quality Audio Signals. An Introduction to the MASCAM System," *EBU Rev.—Tech.*, no. 230, pp. 158–181 (1988 Aug.).

[20] H. Fastl, "Temporal Masking Effects: II. Critical Band Noise Masker," *Acustica*, vol. 36, p. 317 (1977).

[21] K. Brandenburg, "High-Quality Sound Coding at 2.5 Bits/Sample," presented at the 84th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 36, p. 382 (1988 May), preprint 2582.

[22] B. Edler, "Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Functions" (in German), *Frequenz*, vol. 43, pp. 252–256 (1989).

THE AUTHORS

K. Brandenburg

G. Stoll

Karlheinz Brandenburg was born in Erlangen, Germany, in 1954. He received M.S. (Diplom) degrees in electrical engineering in 1980 and in mathematics in 1982 from Erlangen University. In 1989 he earned his Ph.D. in electrical engineering, also from Erlangen University, for work on digital audio coding and perceptual measurement techniques.

From 1989 to 1990 he was with AT&T Bell Laboratories in Murray Hill, NJ, USA. He worked on the ASPEC perceptual coding technique and on the definition of the ISO/IEC MPEG/Audio Layer III system. In 1990 he returned to Erlangen University to continue the research on audio coding and to teach a course on digital audio technology. Since 1993 he has been head of the Audio/Multimedia Department at the Fraunhofer Institute for Integrated Circuits (FhG-IIS).

Dr. Brandenburg has presented numerous papers at AES conventions. In 1994 he received the AES Fellowship Award for his work on perceptual audio coding and psychoacoustic. He is a member of the AES and of the technical committee on Audio and Electroacoustics of the IEEE Signal Processing Society. He is an active member of the ISO MPEG standardization committee, working on advanced audio coding systems. Dr. Brandenburg has been granted 8 patents and has several more pending.

•

Gerhard Stoll studied communications theory and cybernetics at the universities of Stuttgart and München in Germany. After his graduation studies he worked for the Institute of Electroacoustics at the Technical University of München in psychoacoustic research. In 1984, he joined the IRT, a research institute of the German, Swiss, and Austrian broadcasters, where he developed the MUSICAM audio coding system, recently standardized as ISO/MPEG Layer II. Since 1988, he is the head of the group dealing with psychoacoustics and digital audio processing. He is involved in the EUREKA 147 DAB research project, as well as in the international standardization of high-quality audio coding and digital audio broadcasting, e.g. in ISO/IEC MPEG, ETSI, ITU-R and EBU. In 1992, Mr. Stoll received together with Günther Theile and Martin Link the Prof. Lothar Cremer award of the German Acoustical Society for his work with ISO/MPEG Layer II.